ECON 4130
Sept. 05

Lecture Notes to Rice Chapter 5

By H. Goldstein

1.1
Chapter 5 gives an introduction to probabilistic approximation methods, but is insufficient for the needs of an adequate study of econometrics. The commonly non-linear nature of economic models often requires approximation methods for a tractable empirical analysis.

There are many probabilistic convergence concepts available, of which two, *convergence in probability* and *convergence in distribution* are discussed or implied in Rice.

Def. 1 Convergence in Probability.

> *Let* $Y_1, Y_2, \ldots, Y_n, \ldots$ *be a sequence of r.v.'s. Then* Y_n *converges in probability to a constant, c, as* $n \to \infty$ *(written shortly* $Y_n \underset{n\to\infty}{\overset{P}{\to}} c$ *or* $\underset{n\to\infty}{\text{plim}}\, Y_n = c$ *), if , for any* $\varepsilon > 0$, $P(|Y_n - c| > \varepsilon) \to 0$ *as* $n \to \infty$.
> *Or equivalently:* $Y_n \underset{n\to\infty}{\overset{P}{\to}} c$ *if, for any* $\varepsilon > 0$, $P(|Y_n - c| \le \varepsilon) \to 1$ *as* $n \to \infty$.

Example 1: If $X_1, X_2, \ldots$ are *iid* with $E(X_i) = \mu$ and $\text{Var}(X_i) = \sigma^2$, then
$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \underset{n\to\infty}{\overset{P}{\to}} \mu$ (one of the laws of large numbers proven by Chebyshev's inequality).
$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2 \underset{n\to\infty}{\overset{P}{\to}} \sigma^2$, and also $S = \sqrt{S^2} \overset{P}{\to} \sigma$ (proven by the continuity properties of limits in probability described below).

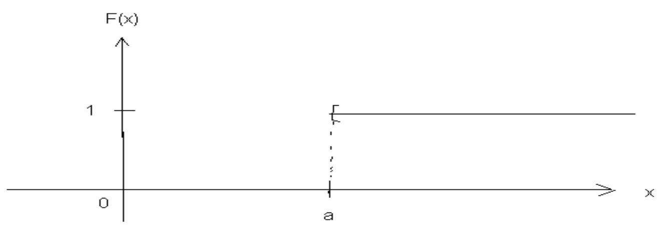
> [**Note on the law of large numbers.** In the lectures we gave a simple proof of the law of large numbers based on Chebyshev's inequality. That proof assumes that the variance, $\text{var}(X_i) = \sigma^2$, exists. It can be proven, however, that this assumption is not necessary. Thus: If $X_1, X_2, \ldots$ are *iid* with $E(X_i) = \mu$, then $\bar{X} \underset{n\to\infty}{\overset{P}{\to}} \mu$, which is a classical result in probability theory.]

1.2 Trivial distributions. It is sometimes convenient to interpret constants as special r.v.'s. Let a be any constant (a real number). We may interpret a as a random variable by introducing the r.v., *X*, by $P(X = a) = 1$. Hence *X* can only take one value (*a*). The probability mass function is then given by $p(a) = P(X = a) = 1$. By the definition of expectation and variance, we have (check formally!), $E(X) = a$ and $\text{var}(X) = 0$.

The cdf of *X* becomes

(1) $$F(x) = P(X \le x) = \begin{cases} 0 & \text{for } x < a \\ 1 & \text{for } x \ge a \end{cases}$$ (see figure 1)

Figure 1



We may call this distribution *the trivial distribution at a*.
Note that $F(x)$ is continuous everywhere except for $x = a$.

(2) The moment generating function (mgf) for *X* is $M(t) = e^{ta}$
(i.e. $M(t) = Ee^{tX} = e^{ta}P(X = a) = e^{ta}$).

Let $a_1, a_2, \ldots, a_n, \ldots$ be a sequence of constants converging to a (in the usual sense) as $n \to \infty$. This means (slightly more precise than given in Sydsæter I): F*or any* $\varepsilon > 0$, *there is a number N such that* $|a_n - a| \le \varepsilon$ *for every* $n \ge N$. From this definition it follows that convergence of sequences in the usual sense can be considered as a special case of convergence in probability.

(3) $\boxed{\textit{If } a_n \underset{n\to\infty}{\to} a, \textit{ then } a_n \overset{P}{\underset{n\to\infty}{\to}} a \quad \textit{(where the } a_n \textit{'s are interpreted as r.v.'s)}}$

Proof: Let $\varepsilon > 0$ be arbitrarily small. We need to show that $P(|a_n - a| \le \varepsilon) \underset{n\to\infty}{\to} 1$. But this probability must be either 0 or 1 according to if $|a_n - a| \le \varepsilon$ is false or true (since $a_n, a,$ and ε are constants and therefore fixed and not subject to random variation). Hence, choosing N such that $|a_n - a| \le \varepsilon$ for all $n \ge N$, we have

$$P(|a_n - a| \le \varepsilon) = \begin{cases} 1 \quad \text{if} \quad |a_n - a| \le \varepsilon \ \text{ is true, which it is for all } \ n \ge N \\ 0 \quad \text{if} \quad |a_n - a| \le \varepsilon \ \text{ is false} \end{cases}$$

This shows that $P(|a_n - a| \le \varepsilon) \underset{n\to\infty}{\to} 1$ since the probability is 1 for all *n* large enough. Q.E.D.

1.3 The continuity property of probability limits.

Theorem 1

(4)

> *Let $X_n, Y_n, \quad n = 1, 2, \ldots$ be two sequences of r.v.'s such that $X_n \overset{P}{\underset{n\to\infty}{\to}} c$ and $Y_n \overset{P}{\underset{n\to\infty}{\to}} d$. Let $g(x)$ be continuous at $x = c$ and $h(x, y)$ be continuous at $x = c$ and $y = d$. Then*
>
> $$g(X_n) \overset{P}{\underset{n\to\infty}{\to}} g(c) \quad \text{and} \quad h(X_n, Y_n) \overset{P}{\underset{n\to\infty}{\to}} h(c, d)$$
>
> *(This is also true when h has more than two arguments.)*

[A proof for those interested is given in **appendix 2**.]

Example 2. Suppose that $X_n \overset{P}{\underset{n\to\infty}{\to}} c$. Then also $Z_n = X_n\left(1 - \dfrac{1}{n}\right) \overset{P}{\underset{n\to\infty}{\to}} c$. Here we use that $h(x, y) = xy$ is continuous, and that $Y_n = 1 - \dfrac{1}{n} \overset{P}{\underset{n\to\infty}{\to}} 1$ because of (3).

Example 3. Suppose that $X_1, X_2, \ldots$ are *iid* with $E(X_i) = \mu$ and $\text{var}(X_i) = \sigma^2$. Then $S^2 = \dfrac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2 \overset{P}{\underset{n\to\infty}{\to}} \sigma^2$ (i.e. S^2 is consistent for σ^2).

Reason: We have $S^2 = \dfrac{n}{n-1}\left[\dfrac{1}{n}\sum_{i=1}^{n} X_i^2 - (\bar{X})^2\right] \overset{P}{\underset{n\to\infty}{\to}} 1 \cdot \left[\mu^2 + \sigma^2 - \mu^2\right] = \sigma^2$

using that, by the law of large numbers (see the note to example 1), $\frac{1}{n}\sum_{i=1}^{n} X_i^{\,2} \xrightarrow{P} E(X_i^{\,2}) = \mu^2 + \sigma^2$, and $\bar{X} \xrightarrow{P} \mu$. Then, use (4) and that $h(x,y) = x - y^2$ is continuous. Finally, use (3) as in example 1. We also obtain that $S = \sqrt{S^2} \xrightarrow{P} \sigma$ since $g(x) = \sqrt{x}$ is continuous.

Exercise 1. Show that the sample correlation, $r = \frac{S_{XY}}{S_X S_Y}$ is a consistent estimator for the population correlation, $\rho = \text{corre}(X,Y) = \frac{\text{cov}(X,Y)}{\sqrt{\text{var}(X)\text{var}(Y)}}$, based on a random sample, $(X_1,Y_1),(X_2,Y_2),\ldots,(X_n,Y_n)$ (meaning that the *n* pairs are independent and have all the same joint distribution). **Hint:** To prove the consistency of the sample covariance, write $S_{XY} = \frac{n}{n-1}\left[\frac{1}{n}\sum_{i=1}^{n} X_i Y_i - \bar{X}\bar{Y}\right]$.

1.4 Convergence in distribution

In the introductory statistics course, the following version of the central limit theorem (CLT) is presented:

Let $X_1, X_2, \ldots$ be iid with $\text{E}(X_i) = \mu$ and $\text{var}(X_i) = \sigma^2$ (implying that $\text{E}(\bar{X}) = \mu$ and $\text{var}(\bar{X}) = \frac{\sigma^2}{n}$). Then, for large *n* ($n \geq 30$ usually considered sufficient), we have

$$Z_n = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} = \frac{\bar{X}-\mu}{\sigma}\sqrt{n} \overset{\text{approximately}}{\sim} N(0,1) \qquad (\text{“}\sim\text{” means “is distributed as”})$$

This statement is somewhat un-precise. What we mean is that “Z_n converges in distribution to *Z*, where *Z* ~ N(0, 1), as $n \to \infty$ ”. (We write this shortly, $Z_n \underset{n\to\infty}{\xrightarrow{D}} Z$, or simply $Z_n \xrightarrow{D} Z$). The formal mathematical definition, given in Rice, is:

Def. 2 (Convergence in distribution)

> *Let $Y_1, Y_2, \ldots$ be a sequence of r.v..'s with cdf's, $F_n(y) = P(Y_n \leq y)$, and Y a r.v. with cdf $F(y) = P(Y \leq y)$. We say that $Y_n \underset{n\to\infty}{\xrightarrow{D}} Y$ if $F_n(y) \underset{n\to\infty}{\to} F(y)$ for every y where the limit cdf, $F(y)$, is continuous.*
>
> (Then, for large *n*, $Y_n \overset{\text{approx.}}{\sim} F(y)$)

This means: If the limit cdf, $F(y)$, is continuous for $y = a$ and $y = b$, then
$P(a < Y_n \le b) = F_n(b) - F_n(a) \underset{n\to\infty}{\to} F(b) - F(a) = P(a < Y \le b)$
Hence, $P(a < Y_n \le b) \approx P(a < Y \le b)$ for large *n*.

Note that, if the limit distribution is N(0,1) (which is most often the case), then the limit cdf (usually written $\Phi(x) = P(Z \le x)$ where $Z \sim N(0,1)$) is continuous for all *x*.

Another useful comment is that *convergence in probability* can be interpreted as a special case of *convergence in distribution* by the following lemma:

(5)

> $Y_n \underset{n\to\infty}{\overset{P}{\to}} c$ *is equivalent to* $Y_n \underset{n\to\infty}{\overset{D}{\to}} Y$ *where Y is the trivial r.v. at c (i.e.* $P(Y = c) = 1$*) with the trivial cdf as in* (1). *(The last statement we may simply write* $Y_n \underset{n\to\infty}{\overset{D}{\to}} c$ *.)*

[For those interested, a proof is written out in appendix 2.]

1.5 Determination of limit distributions

It turns out difficult (usually) to use the definition of limit in distribution directly to derive a limit distribution. Therefore, there has been developed a number of techniques and tools in the literature for this purpose. One important tool is by means of moment generating functions (mgf's) formulated as theorem A in Rice, chapter 5, and cited below in theorem 2. (An even more important tool is by means of so-called *characteristic functions*, (see Rice at the end of section 4.5), which requires complex analysis and is omitted here.)

Theorem 2 (Theorem A in Rice, chapter 5)

> *Let* Y_n, $n = 1, 2, \ldots$ *be a sequence of r.v.'s with cdf's,* $Y_n \sim F_n(y) = P(Y_n \le y)$. *Suppose that the mgf's,* $M_n(t) = Ee^{tY_n}$ *, exist for all n. Let Y be a r.v. with cdf,* $F(y)$ *and mgf* $M(t) = Ee^{tY}$ *, and assume that* $M_n(t) \underset{n\to\infty}{\to} M(t)$ *for all t in an open interval that contains 0. Then* $Y_n \underset{n\to\infty}{\overset{D}{\to}} Y$ *(i.e.* $F_n(y) \underset{n\to\infty}{\to} F(y)$ *for all y where* $F(y)$ *is continuous).*

Example 4 (example A in Rice, section 5.3)

We simplify the argument in Rice by using l'Hôpital's rule instead of his series argument.
Let $X_n \sim \text{pois}(\lambda_n),\ n = 1, 2, \ldots$ where $\lambda_1, \lambda_2, \ldots$ is a sequence of numbers such that $\lambda_n \underset{n\to\infty}{\to} \infty$. Then $\mathrm{E}(X_n) = \text{var}(X_n) = \lambda_n$. We will show that the standardized

$$Z_n = \frac{X_n - \mathrm{E}(X_n)}{\sqrt{\text{var}(X_n)}} = \frac{X_n - \lambda_n}{\sqrt{\lambda_n}} = \frac{1}{\sqrt{\lambda_n}} X_n - \sqrt{\lambda_n}$$

converges in distribution to $Z \sim N(0,1)$, which follows if we can show that the *mgf* of Z_n converges to the *mgf* of $Z \sim N(0, 1)$, i.e. $M_Z(t) = e^{t^2/2}$. The mgf of X_n is (see Rice, section 4.5, example A):

$$M_{X_n}(t) = e^{\lambda_n (e^t - 1)}$$

We have from before that, if *X* and *Y* are r.v.'s such that $Y = a + bX$, the *mgf* of Y is, $M_Y(t) = e^{at} M_X(bt)$. Hence

$$M_{Z_n}(t) = e^{-t\sqrt{\lambda_n}} M_{X_n}\left(\frac{1}{\sqrt{\lambda_n}} t\right) = e^{-t\sqrt{\lambda_n}} \cdot e^{\lambda_n (e^{t/\sqrt{\lambda_n}} - 1)} \qquad \text{or}$$

$$\ln(M_{Z_n}(t)) = -t\sqrt{\lambda_n} + \lambda_n \left(e^{t/\sqrt{\lambda_n}} - 1\right) \quad \text{(notice printing mistake in Rice)}$$

Put $x = \frac{1}{\sqrt{\lambda_n}}$. Since $\lambda_n \underset{n\to\infty}{\to} \infty$, we have $x \underset{n\to\infty}{\to} 0$. From l'Hôpital's rule we get

$$\ln(M_{Z_n}(t)) = -\frac{t}{x} + \frac{1}{x^2}\left(e^{xt} - 1\right) = \frac{e^{xt} - 1 - xt}{x^2} \underset{x\to 0}{\to} \lim_{x\to 0} \frac{te^{xt} - t}{2x} = \lim_{x\to 0} \frac{t^2 e^{tx}}{2} = \frac{t^2}{2}$$

Since e^x is a continuous function of *x*, $M_{Z_n}(t) \underset{n\to\infty}{\to} e^{t^2/2}$, implying $Z_n \underset{n\to\infty}{\overset{D}{\to}} Z \sim N(0,1)$.
(*End of example.*)

We will now repeat Rice's proof of the central limit theorem (CLT) supplied with some details.

Theorem 3 (CLT, theorem B in Rice, section 5.3)

Let $X_1, X_2, \ldots$ *be a sequence of iid r.v.'s with* $\mathrm{E}(X_i) = 0$ *and* $\mathrm{var}(X_i) = \sigma^2$.
Let $S_n = \sum_{i=1}^{n} X_i$ *(implying* $\mathrm{E}(S_n) = 0$ *and* $\mathrm{var}(S_n) = n\sigma^2$*). Then*
$\frac{S_n}{\sqrt{\mathrm{var}(S_n)}} = \frac{S_n}{\sigma\sqrt{n}} \underset{n\to\infty}{\overset{D}{\to}} Z \sim N(0,1)$ *(or* $P\left(\frac{S_n}{\sigma\sqrt{n}} \le x\right) \underset{n\to\infty}{\to} \Phi(x)$ *for all x since*
$\Phi(x) = P(Z \le x)$ *is continuous everywhere).*

[**Note.** The proof is only given here for the special case that the *mgf* of X_j, $M(t) = \mathrm{E}(e^{tX_j})$, exists in an open interval containing 0, which is not always the case (see the note to (A5) in appendix 1). The proof for the general case is almost identical to the given one, but based instead on characteristic functions (defined by $g(t) = \mathrm{E}(e^{itX_j})$ where *i* is the complex number, $\sqrt{-1}$). Characteristic functions exist for every probability distribution. Such a proof, however, requires some knowledge of complex analysis, and is omitted here.]

Proof. Assume that the common *mgf* of $X_1, X_2, \ldots$, $M(t) = \mathrm{E}(e^{tX_i})$, exists in an open interval, (a,b), where $a < 0 < b$. Then, according to (A5) in appendix 1, $M(t)$, has continuous derivatives of all orders in (a,b).

Since $X_1, X_2, \ldots$ are independent and identically distributed, the *mgf* of S_n is

$$M_{S_n}(t) = \mathrm{E}\left(\mathrm{e}^{t\sum_{i=1}^{n} X_i}\right) = \mathrm{E}\left(e^{tX_1} e^{tX_2} \cdots e^{tX_n}\right) = \mathrm{E}\left(e^{tX_1}\right)\mathrm{E}\left(e^{tX_2}\right)\cdots\mathrm{E}\left(e^{tX_n}\right) = M(t)^n$$

Putting $Z_n = \frac{S_n}{\sigma\sqrt{n}}$, we obtain the *mgf*, $M_{Z_n}(t) = M\left(\frac{t}{\sigma\sqrt{n}}\right)^n$
Applying Taylor's formula (see (A2) in appendix 1) to $M(t)$, we have

$M(t) = M(0) + tM'(0) + \frac{t^2}{2}M''(0) + \frac{t^3}{3!}M'''(c)$ where c is somewhere between 0 and t. We have $M(0) = \mathrm{E}(e^{0X_i}) = 1$, $M'(0) = \mathrm{E}(X_i) = 0$, and $M''(0) = \mathrm{E}(X_i^2) = \sigma^2$. Hence

$$M(t) = 1 + \frac{t^2}{2}\sigma^2 + \frac{t^3}{6}M'''(c)$$

Substituting into $M_{Z_n}(t)$, we obtain

$$M_{Z_n}(t) = M\left(\frac{t}{\sigma\sqrt{n}}\right)^n = \left[1 + \frac{\left(\frac{t}{\sigma\sqrt{n}}\right)^2}{2}\sigma^2 + \frac{\left(\frac{t}{\sigma\sqrt{n}}\right)^3}{6} M'''(c_n)\right]^n$$

or

$$M_{Z_n}(t) = \left[1 + \frac{t^2}{2n} + R_n\right]^n$$ where $R_n = \frac{t^3}{6\sigma^3 n^{\frac{3}{2}}} M'''(c_n)$, and c_n lies between 0 and $\frac{t}{\sigma\sqrt{n}}$.

We will now prove that $n \cdot R_n \underset{n\to\infty}{\to} 0$, i.e. $n \cdot R_n = \frac{t^3}{6\sigma^3\sqrt{n}} M'''(c_n) \underset{n\to\infty}{\to} 0$

Since c_n lies between 0 and $\frac{t}{\sigma\sqrt{n}}$, and $\frac{t}{\sigma\sqrt{n}} \underset{n\to\infty}{\to} 0$, we must have that $c_n \to 0$. Therefore, $M'''(c_n) \to M'''(0)$ since $M'''(t)$ is continuous in 0 (see (A5) in appendix 1). Hence, $M'''(c_n)$ is bounded, and $M'''(c_n)/\sqrt{n} \to 0$, which proves that $n \cdot R_n \underset{n\to\infty}{\to} 0$.

We finally get $M_{Z_n}(t) = \left[1 + \frac{t^2}{2n} + R_n\right]^n = \left[1 + \frac{a_n}{n}\right]^n$ where $a_n = \frac{t^2}{2} + n \cdot R_n \to \frac{t^2}{2}$

Thus, using (A3) in appendix 1, we get $M_{Z_n}(t) \to e^{t^2/2}$, which is the *mgf* of $N(0,1)$. Property A in Rice, section 4.5, tells us that the *mgf* uniquely determines the probability distribution. Hence, $Z_n \underset{n\to\infty}{\overset{D}{\to}} Z \sim N(0,1)$. Q.E.D.

In practice the following reformulation of the CLT is the most common:

Corollary (CLT)

> *If* $X_1, X_2, \ldots \sim$ *iid, with* $\mathrm{E}(X_i) = \mu$ *and* $\mathrm{var}(X_i) = \sigma^2$, *then*
> $\frac{\sqrt{n}(\bar{X} - \mu)}{\sigma} \underset{n\to\infty}{\overset{D}{\to}} Z \sim N(0,1)$, *which means that* $\bar{X} \overset{\text{approximately}}{\sim} N(\mu, \frac{\sigma^2}{n})$
> *for large n.*

Proof. Put $Y_i = X_i - \mu$. Then, $Y_1, Y_2, \ldots \sim iid$, $\mathrm{E}(Y_i) = 0$ and $\mathrm{var}(Y_i) = \sigma^2$, and we can use theorem 3:

$$\frac{\sum_{i=1}^{n} Y_i}{\sigma\sqrt{n}} = \frac{n\bar{X} - n\mu}{\sigma\sqrt{n}} = \frac{\sqrt{n}(\bar{X} - \mu)}{\sigma} \underset{n\to\infty}{\overset{D}{\to}} Z \sim N(0,1)$$

$$\Rightarrow \frac{\sqrt{n}(\bar{X}-\mu)}{\sigma} \overset{\text{approx.}}{\sim} N(0,1) \text{ for large } n \Rightarrow \sqrt{n}(\bar{X}-\mu) \overset{\text{approx.}}{\sim} N(0,\sigma^2) \text{ for large } n$$

$$\Rightarrow \bar{X} - \mu \overset{\text{approx.}}{\sim} N(0, \frac{\sigma^2}{n}) \text{ for large } n \Rightarrow \bar{X} \overset{\text{approx.}}{\sim} N(\mu, \frac{\sigma^2}{n}) \text{ for large } n. \quad \text{Q.E.D.}$$

The last result we present, is an extremely useful result for statistical practice:

Theorem 4 (Slutsky's lemma)

Let A_n, B_n, X_n *be r.v.'s such that* $A_n \underset{n\to\infty}{\overset{P}{\to}} a$ *(constant),* $B_n \underset{n\to\infty}{\overset{P}{\to}} b$ *(constant),*

and $X_n \underset{n\to\infty}{\overset{D}{\to}} X$. *Then* $A_n X_n + B_n \underset{n\to\infty}{\overset{D}{\to}} aX + b$

In particular, if $A_n \underset{n\to\infty}{\overset{P}{\to}} 0$, *then* $A_n X_n + B_n \underset{n\to\infty}{\overset{P}{\to}} b$ *(because of (5) above).*

The proof is a straightforward, but somewhat lengthy, ε,δ - argument along the lines illustrated in appendix 2, and is omitted here.

Here we illustrate the result by making some arguments for confidence intervals presented in the introductory statistics course more precise.

Example 5. (Confidence intervals)
Suppose $X_1, X_2, \ldots$ are *iid*, with $\mathrm{E}(X_i) = \mu$ (unknown) and $\mathrm{var}(X_i) = \sigma^2$. We want a confidence interval (CI) with degree of confidence, $1-\alpha$, for the unknown μ. Even if the common distribution, $F(x)$, for the X_i's, is unknown, the distribution of $\bar{X}$ is approximately known for large *n* ($n \ge 30$ usually considered sufficient) because of the CLT, which we utilize as follows.

For large *n*, $Z_n = \dfrac{\bar{X}-\mu}{\sigma/\sqrt{n}} \overset{\text{approx.}}{\sim} N(0,1)$. Hence, $P(-z_{\frac{\alpha}{2}} \le Z_n \le z_{\frac{\alpha}{2}}) \approx 1-\alpha$ where $z_{\frac{\alpha}{2}}$ is the upper $\alpha/2$-point in $N(0,1)$. Manipulating the probability, we get

$$P\left(\bar{X} - z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \le \mu \le \bar{X} + z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \right) \approx 1-\alpha$$

Thus, if σ is known, then an approximately $1-\alpha$ CI for μ is given by

$$\bar{X} \pm z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$$

In practice σ is usually unknown, but according to Slutzky's lemma, can be replaced by a consistent estimator, as the following argument shows:

Put $U_n = \dfrac{\bar{X}-\mu}{\hat{\sigma}\sqrt{n}}$ where $\hat{\sigma} = \sqrt{S^2} = \sqrt{\dfrac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2}$ is consistent for σ (see example 3). We then have

$$U_n = \frac{\bar{X}-\mu}{\hat{\sigma}\sqrt{n}} = \frac{\sigma}{\hat{\sigma}} \cdot \frac{\bar{X}-\mu}{\sigma\sqrt{n}} = \frac{\sigma}{\hat{\sigma}} \cdot Z_n$$

Since $\dfrac{\sigma}{\hat{\sigma}} \underset{n\to\infty}{\overset{P}{\to}} \dfrac{\sigma}{\sigma} = 1$ (see theorem 1 and example 3), we have from Slutzky's lemma $U_n \underset{n\to\infty}{\overset{D}{\to}} 1 \cdot Z = Z \sim N(0,1)$. Hence, for large *n*, $P(-z_{\frac{\alpha}{2}} \le U_n \le z_{\frac{\alpha}{2}}) \approx 1-\alpha$. Manipulating this, we get

$$P\left(\bar{X} - z_{\frac{\alpha}{2}} \frac{\hat{\sigma}}{\sqrt{n}} \le \mu \le \bar{X} + z_{\frac{\alpha}{2}} \frac{\hat{\sigma}}{\sqrt{n}} \right) \approx 1-\alpha$$

which gives the approximate $1-\alpha$ CI for μ: $\bar{X} \pm z_{\frac{\alpha}{2}} \dfrac{\hat{\sigma}}{\sqrt{n}}$. Simulation studies show that the approximation is usually satisfactory for $n \ge 30$.

We have a similar state of affairs for poisson- and binomial models:

The poisson case: Suppose that the number, *X*, of working accidents during *t* time units in a large firm, is ~ pois(λt), where λ is the unknown expected (i.e. long run average) accident rate per time unit in the firm. Then $\mathrm{E}(X) = \lambda t = \mathrm{var}(X)$, which implies that $\hat{\lambda} = \dfrac{X}{t}$ is an unbiased estimator of λ. Since $\mathrm{var}(\hat{\lambda}) = \dfrac{\lambda}{t} \underset{t\to\infty}{\to} 0$, it follows from Chebyshev's inequality (check!) that $\hat{\lambda}$ is consistent for λ as well (i.e., $\hat{\lambda} \underset{t\to\infty}{\overset{P}{\to}} \lambda$). From example 4 we get that

$$Z_t = \frac{X - t\lambda}{\sqrt{t\lambda}} = \frac{t\hat{\lambda} - t\lambda}{\sqrt{t}\sqrt{\lambda}} = \sqrt{t}\,\frac{\hat{\lambda} - \lambda}{\sqrt{\lambda}} \underset{t\to\infty}{\overset{D}{\to}} Z \sim N(0,1) \quad \text{since } \ t\lambda \to \infty \ \text{ as } \ t \to \infty .$$

Slutzky's lemma shows that we can replace λ by $\hat{\lambda}$ in the denominator of Z_t without destroying the approximation substantially, i.e.,

$$U_t = \sqrt{t}\,\frac{\hat{\lambda} - \lambda}{\sqrt{\hat{\lambda}}} = \frac{\sqrt{\lambda}}{\sqrt{\hat{\lambda}}} \cdot Z_t \underset{t\to\infty}{\overset{D}{\to}} 1 \cdot Z = Z \sim N(0,1) \quad \text{since } \ \frac{\sqrt{\lambda}}{\sqrt{\hat{\lambda}}} \overset{P}{\to} 1 \ \text{ as } \ t \to \infty \text{, using}$$

that the function, $g(x) = \sqrt{\lambda}\big/\sqrt{x}$ is continuous in *x*. We then get for large *t* (the criterion $t\lambda \geq 10$ is usually considered sufficient), the following approximation

$$P\left(\hat{\lambda} - z_{\frac{\alpha}{2}} \frac{\sqrt{\hat{\lambda}}}{\sqrt{t}} \leq \lambda \leq \hat{\lambda} + z_{\frac{\alpha}{2}} \frac{\sqrt{\hat{\lambda}}}{\sqrt{t}} \right) \approx 1 - \alpha$$

giving an approximate $1-\alpha$ CI for λ: $\hat{\lambda} \pm z_{\frac{\alpha}{2}} \dfrac{\sqrt{\hat{\lambda}}}{\sqrt{t}}$.

Discuss the binomial case yourself.

Appendix 1 (mathematical prerequisites for Rice, chapter 5)

The students are recommended to read Sydsæter I, section 6.4 on sequences ("tallfølger") and section 7.6 on Taylor polynomials and series.

The following result is much used in probability theory (see a motivation in Sydsæter I, section 7.6).

(A1)

For any real a, e^a can be expressed as an infinite series

$$e^a = \sum_{i=0}^{\infty} \frac{a^i}{i!} = 1 + a + \frac{a^2}{2!} + \cdots + \frac{a^n}{n!} + \cdots$$

If c is a common factor, it can be taken outside the sum,

$$\sum_{i=0}^{\infty} c\,\frac{a^i}{i!} = c \sum_{i=0}^{\infty} \frac{a^i}{i!} = ce^a$$

[**Note.** The theory of infinite series is not treated in the mathematics curriculum, except geometric series, so we will not go into this here. We only mention that the precise mathematical meaning of the sum is a limit of a sequence of numbers (see Sydsæter I, section 6.4 for the meaning of a sequence):

$$e^a = \lim_{n\to\infty}\left(1 + a + \frac{a^2}{2!} + \cdots + \frac{a^n}{n!}\right),$$

which can be shown to be well defined (we say that the series is *convergent*) for every *a*. The only result from the theory of infinite series we use is the last statement that a common factor can be taken outside the sum. The series is mainly used in this course to derive the *mgf* for a poisson r.v. (see Example A in Rice, section 4.5): $X \sim \text{pois}(\lambda)$ implies that the *mgf* is $M(t) = \mathrm{E}(e^{tX}) = e^{\lambda(e^t - 1)}$]

Much of approximation theory in mathematics and probability theory is based on the famous **Taylor's formula**, given in **(A2)** (see Sydsæter I, section 7.6):

(A2)

Let $f(x)$ be $n+1$ times differentiable in an interval that contains 0 and x. Then, $f(x)$ can be approximated by a polynomial as follows

$$f(x) = f(0) + \frac{x}{1!}f'(0) + \frac{x^2}{2!}f''(0) + \cdots + \frac{x^n}{n!}f^{(n)}(0) + R_{n+1}(x)$$

where the error term, $R_{n+1}(x)$, is $R_{n+1}(x) = \frac{x^{n+1}}{(n+1)!}f^{(n+1)}(c)$, where c is a number lying somewhere between 0 and x.

[**Note:** In **(A2)** we say that $f(x)$ is expanded around $x = 0$. From **(A2)** it follows that we can expand $f(x)$ around any other value, $x = \mu$, where *f* is differentiable: Write $f(x) = f(\mu + x - \mu)$ and define $g(h) = f(\mu + h)$ where $h = x - \mu$. Then $g(0) = f(\mu)$ and $g^{(n)}(0) = f^{(n)}(\mu)$. Applying **(A2)** to $g(h)$, we obtain an expansion of $f(x)$ around $x = \mu$:

(A3) $$f(x) = g(x-\mu) = f(\mu) + \frac{x-\mu}{1!}f'(\mu) + \cdots + \frac{(x-\mu)^n}{n!}f^{(n)}(\mu) + \frac{(x-\mu)^{n+1}}{(n+1)!}f^{(n+1)}(c)$$

where c is a number lying somewhere between μ and *x*.]

Example 6. Rice section 4.6 gives examples of finding approximate expressions of expectations and variances. Let *X* be a r.v. with $\mathrm{E}(X) = \mu$ and $\mathrm{var}(X) = \sigma^2$. Suppose we want the expectation and variance of a transformed r.v., $Y = g(X)$. If *g* is complicated it is often hard to find $E(Y)$ and var(*Y*) exactly. If $g(x)$ is differentiable around $x = \mu$, however, we can easily obtain approximate values by using Taylor expansion around μ. Ignoring the error term, we have from **(A3)** with *n* =1:

$$g(X) \approx g(\mu) + g'(\mu)(X - \mu)$$

By taking expected value and variance on both sides, we get (note that $g(\mu)$ and $g'(\mu)$ are constants)

$$E(g(X)) \approx g(\mu) \quad \text{and} \quad \text{var}(g(X)) \approx [g'(\mu)]^2 \sigma^2$$

By including an extra term in the expansion, we may obtain a (hopefully – it depends on the error term) better approximation to the expectation:

$$g(X) \approx g(\mu) + g'(\mu)(X-\mu) + \frac{g''(\mu)}{2}(X-\mu)^2$$

gives

$$E(g(X)) \approx g(\mu) + \frac{\sigma^2}{2} g''(\mu) \text{ (read example B in Rice, sec. 4.6)}$$

Note that it is usually not a good idea in this context to include many terms in the Taylor approximation since terms like $(X-\mu)^r$ for larger *r* are often statistically quite unstable, which may destroy the approximation. (*End of example.*)

From **(A2)** we can now derive the following much used result (also used in the proof of the CLT):

(A4) *If* $a_n, \quad n = 1, 2, \ldots$ *is a sequence of numbers(see Sydsæter I, section 6.4) converging to a number, a (i.e.* $a_n \underset{n\to\infty}{\longrightarrow} a$ *), then*

$$\left(1 + \frac{a_n}{n}\right)^n \underset{n\to\infty}{\longrightarrow} e^a$$

Proof: The result follows if we can show that $n \cdot \ln\left(1 + \frac{a_n}{n}\right) \underset{n\to\infty}{\longrightarrow} a$ (since e^x is a continuous function).

Put $x_n = \frac{a_n}{n}$. Then $n \cdot x_n = a_n \underset{n\to\infty}{\longrightarrow} a$. Applying **(A2)** to the function, $f(x) = n \cdot \ln(1+x)$, with only one term plus error, we get $f(x) = f(0) + f'(c)x = \frac{n}{1+c}x$, where *c* is between 0 and *x*. Note that $f(0) = 0$. Therefore, $f(x_n) = n \cdot \ln(1 + x_n) = \frac{n \cdot x_n}{1 + c_n} \underset{n\to\infty}{\longrightarrow} \frac{a}{1} = a$, using that $n \cdot x_n \underset{n\to\infty}{\longrightarrow} a$ and that $c_n \underset{n\to\infty}{\longrightarrow} 0$. The last statement follows since c_n always lies between 0 and x_n (implying $0 \le |c_n| \le |x_n|$), and $x_n = \frac{a_n}{n} \underset{n\to\infty}{\longrightarrow} 0$ since the sequence, $a_n, \; n = 1, 2, \ldots$ converges to *a*, and therefore must be bounded (i.e., there is a number *C* such that $|a_n| \le C$ for all *n*).
Q.E.D.

In order to make the proof of the CLT completely rigorous we need one more mathematical fact.

(A5) *If the mgf, $M(t) = \mathrm{E}(e^{tX})$ of a r.v., X, exists for all t in an open interval containing 0 (i.e. for all $t \in (a,b)$ where $a < 0 < b$), then the n-th derivative, $M^{(n)}(t)$, exists for all $n = 1, 2, \ldots$ in this interval. This implies, in particular that $M^{(n)}(t)$ is continuous in (a,b) for all n.*

[**Note.** This result is not hard to prove, but requires results from more advanced integration theory, and is therefore omitted here. Note also that **(A5)** shows that the assumption that $M(t)$ exists in an open interval around 0, is a quite strong assumption on the distribution of *X*. It implies that moments, $\mathrm{E}(X^r)$, of all orders $r = 1, 2, \ldots$ exist. This follows since, $\mathrm{E}(X^r) = M^{(r)}(0)$ then exists for all *r*. The assumption is valid for most of the common distributions met in this course, but there are notable exceptions. For example it is not true for *t*-distributions, since, if *X* is *t*-distributed with ν degrees of freedom, then it can be shown that $\mathrm{E}(X^r)$ exists only for $r < \nu$.]

Appendix 2 (some proofs)

Proof of (4) (optional reading)

We will prove the *h*(*x*,*y*)-case. Try to write out a proof for the simpler *g*(*x*)-case yourself (in case you don't realize that the *g*-case follows directly from the *h*-case).

Suppose $X_n \underset{n\to\infty}{\overset{P}{\to}} c$ and $Y_n \underset{n\to\infty}{\overset{P}{\to}} d$ and that $h(x,y)$ is continuous for $x = c,\ y = d$. Choose an $\varepsilon > 0$ arbitrarily small. We need to prove that $P(|\, h(X_n, Y_n) - g(c,d)\,| \le \varepsilon) \underset{n\to\infty}{\to} 1$.

According to the meaning of continuity (see e.g. Sydsæter I, sec. 6.9), there is a $\delta > 0$ such that, whenever $|\, x - c\,| \le \delta$ and $|\, y - d\,| \le \delta$, then $|\, h(x,y) - h(c,d)\,| \le \varepsilon$.

Define events, A_n, B_n, C_n by $A_n = (|\, X_n - c\,| \le \delta)$, $B_n = (|\, Y_n - d\,| \le \delta)$, and $C_n = (|\, h(X_n, Y_n) - h(c,d)\,| \le \varepsilon)$.

We then have $A_n \cap B_n \Rightarrow C_n$ which implies that $P(C_n) \ge P(A_n \cap B_n)$. (Note that if *A*, *B* are events such that $A \Rightarrow B$, or $A \subset B$ interpreted as sets, then $P(A) \le P(B)$).

According to the definition of probability limit, $P(A_n) \underset{n\to\infty}{\to} 1$ and $P(B_n) \underset{n\to\infty}{\to} 1$.

This implies that $P(A_n \cap B_n) \underset{n\to\infty}{\to} 1$ since

$P(A_n \cap B_n) = P(A_n) + P(B_n) - P(A_n \cup B_n) \to 1+1-1=1$ as $n \to \infty$ (Note that $P(A_n \cup B_n) \geq P(A_n) \to 1$ implies that $P(A_n \cup B_n) \to 1$). Hence, since $P(C_n) \geq P(A_n \cap B_n)$, also $P(C_n) \to 1$ as $n \to \infty$. Q.E.D.

Proof of (5) (optional reading)

i) Suppose that $Y_n \underset{n\to\infty}{\overset{P}{\to}} c$. We need to prove that $Y_n \underset{n\to\infty}{\overset{D}{\to}} Y$ where $P(Y=c)=1$. Let the cdf of Y_n be $F_n(y)$ and the cdf of Y be $F(y)$, i.e. the trivial cdf at c (see **1.2**)

$$F(y) = P(Y \leq y) = \begin{cases} 0 & \text{for} \quad y < c \\ 1 & \text{for} \quad y \geq c \end{cases}$$

Thus $F(y)$ is continuous for all $y \neq c$.

Hence, according to the definition of convergence in distribution, we need to show that $F_n(y) \underset{n\to\infty}{\to} F(y)$ for all $y \neq c$, or $F_n(y) \underset{n\to\infty}{\to} 0$ for $y < c$ and $F_n(y) \underset{n\to\infty}{\to} 1$ for $y > c$.

Again we use that if $A \Rightarrow B$, then $P(A) \leq P(B)$. Suppose $y > c$ (or $y - c > 0$). Then the following events satisfy

$$(|Y_n - c| \leq y - c) \Leftrightarrow (-(y-c) \leq Y_n - c \leq y - c) \Leftrightarrow (c - (y-c) \leq Y_n \leq c + y - c)$$
$$\Leftrightarrow (2c - y \leq Y_n \leq y) \Rightarrow (Y_n \leq y)$$

Hence $F_n(y) = P(Y_n \leq y) \geq P(|Y_n - c| \leq y - c) \underset{n\to\infty}{\to} 1$ since $Y_n \underset{n\to\infty}{\overset{P}{\to}} c$. Therefore, we must have that $F_n(y) \underset{n\to\infty}{\to} 1$.

Now, suppose $y < c$ (i.e. $c - y > 0$). We have

$$(Y_n \leq y) \Leftrightarrow (-Y_n \geq -y) \Leftrightarrow (c - Y_n \geq c - y) \Rightarrow (|Y_n - c| \geq c - y) \Rightarrow (|Y_n - c| > \frac{c-y}{2})$$

Thus, $F_n(y) = P(Y_n \leq y) \leq P\left(|Y_n - c| > \dfrac{c-y}{2}\right) \underset{n\to\infty}{\to} 0$, which implies that $F_n(y) \underset{n\to\infty}{\to} 0$, and we have proven that $Y_n \underset{n\to\infty}{\overset{D}{\to}} Y$.

ii) Now, conversely, suppose that $Y_n \underset{n\to\infty}{\overset{D}{\to}} Y$ where $P(Y=c)=1$. Then $F_n(y) \underset{n\to\infty}{\to} F(y)$ for all $y \neq c$. Let $\varepsilon > 0$ be arbitrary small. We have

$$P(|Y_n - c| \leq \varepsilon) = P(c - \varepsilon \leq Y_n \leq c + \varepsilon) \geq P(c - \varepsilon < Y_n \leq c + \varepsilon) = F_n(c+\varepsilon) - F_n(c-\varepsilon)$$

Since $F(y)$ is continuous for $y = c - \varepsilon$ and $y = c + \varepsilon$, the last expression converges to $F(c+\varepsilon) - F(c-\varepsilon) = 1 - 0 = 1$ as $n \to \infty$. Hence $P(|Y_n - c| \le \varepsilon) \underset{n\to\infty}{\to} 1$, and we have proven that $Y_n \overset{P}{\underset{n\to\infty}{\to}} c$. Q.E.D.